Academic Assessment of Limited English Proficient (LEP[1]) Students in the Era of Accountability: A Review of the Literature

Tim Boals, Consultant
Bilingual/ESL Programs

Introduction

Recently, significant changes have occurred in federal law requiring the inclusion of all children in statewide, or large-scale assessment in an effort to increase educational accountability at the district and school levels. The Improving America's Schools Act (IASA) of 1994 and the Individuals with Disabilities Education Act (IDEA), as amended in 1997, both require districts to report to state departments of education data about student progress that are complete and disaggregated by educationally significant categories, including disability groups and limited proficiency in English.

For years, many educational professionals advocated the exclusion of these students from standardized measures of student achievement, maintaining that the assessments were biased and potentially harmful to special needs students who should not be held to the proficiency norms of the general population. While some argue the assessments still raise the same issues of lack of appropriateness or cultural bias, today's educators lament the practice of exclusion from statewide assessment. The shift in thinking regarding the assessment of students with disabilities and limited English proficiency (LEP) is largely due to the realization that students who are left out of educational assessments are too often left out of curricular reforms and program improvement efforts which increase student achievement.

The assessment policies that states adopt directly affect the curricular goals, instructional practices and educational outcomes for these students. While much has been written about the need for accountability in general, very little is specifically known about the impact current assessment policy decisions will have on the students they are intended to serve. Furthermore, little is known about how the individual, state-by-state, variations in response to the federal legislation might impact, either positively or negatively, special needs students.

In the era of accountability, states must grapple with the issue of how to enforce the "All Means All" edict while insuring that assessments are reasonably fair and accurate measures of student progress. For many students, this will simply mean being included when they would not have been in the previous era. For others, it will mean taking the required assessments, but with appropriate accommodations to allow students to better demonstrate what they know and are able to do. For a relatively smaller percentage of students[2], it may mean taking an alternate assessment or series of assessments that are, ideally, aligned with the same curricular goals and academic standards of other students, but presented in formats that respond to the unique needs of these learners. (Zehler et al., 1994; NCBE, 1997)

In this review, I examine the literature on large-scale, academic assessment of LEP students for accountability purposes, focusing within topic area subsections on key issues and concerns. For point of comparison only, issues raised in the more extensive literature into accommodations and alternate assessment for special education students are also considered; however, no conclusions are drawn with respect to how large-scale assessments may impact students with disabilities.

As we look at the literature in this area, it will be useful to consider who is being held accountable for what, and how soon. Particularly, in the case of high stakes testing, we ask whether the assessments are valid for the purposes to which they are asked to serve, and whether the consequences are reasonable and fair. This includes individual student and program performance consequences as large-scale testing increasingly becomes the primary means for demonstrating accountability. A summary of key issues and concerns is offered in the conclusion.

The Role of Language and Culture in Measuring Academic Performance

Perhaps the most obvious and often the first issue to be considered is the LEP student's English language proficiency level. While virtually all authors agree that language proficiency is a crucial issue, there is a surprising lack of agreement on when students actually have reached a proficiency level at which tests scores can be considered valid and reliable measures of student performance. Zehler et al. (1994), for example, states that standardized achievement testing (either traditional or alternative) should not begin prior to one and one-half years of instruction in English. Given the research of Cummins and others who find that academic English skills usually take five to seven years to fully develop, this seems, at first glance, an inadequate time frame for participation in traditional large-scale assessments. (August & Hakuta, 1997)

Most of the literature suggests two or even three years as a reasonable period for students to be exempted from traditional, subject area achievement testing. LaCelle-Peterson and Rivera (1994) maintain, however, that the validity issue is still a very real concern even for students at more intermediate English proficiency levels that may conceivably take four or even five years to attain. While certainly in support of the notion that LEP students, and the programs that serve them, need to be a part of the accountability process, what these authors are saying is that we must be aware of the very real possibility that testing data can, and often is, used to make educational decisions that may adversely affect these same children and programs. This is the real issue that deserves our attention. Including an LEP student in testing after one and one-half years may be fine if we are focusing on monitoring student progress or program gains over time. The minute we begin to use such data to make high stakes decisions like program placements, retention-in-grade, or high school graduation, higher standards of validity are required. Particularly for LEP students and the programs that serve them, there is a legitimate question as to whether those standards are being met. Many educators and policymakers continue to largely miss the point that it is not inclusion or exclusion per se that presents the problem, rather, it is what we do with the results and whether we have adequately taken into consideration the complex factors affecting LEP achievement, to say nothing of the limits of the testing itself.

LaCelle-Peterson and Rivera's (1994) cautionary stance stems, in part, from the potential to misuse testing data. They remind us that "historically, we as U.S. educators have not applied the best educational thinking to issues facing ELLs (English Language Learners)" (p. 57). In the rush to correct for the lack of accountability and inclusion in the past, are we in danger of including students in masse and subsequently using that information in ways that are neither valid nor fair?

This potential seems even more real as we consider other issues treated within the literature. Anderson et al., in an article based on national survey data, (1997) details an array of factors we must keep in mind with respect to the testing of LEP students. These factors include cultural bias of the testing instruments, cultural values of the student, and socioeconomic bias to

name only a few. They further point out that students who are not fully acculturated to middle class mainstream values and beliefs may have differing attitudes towards gender roles, the individual versus the group or family, the use of time, and a host of other factors including a lack of familiarity with the testing format.

Researchers know that language acquisition and acculturation occur at widely differing rates for different students. This is one factor that makes problematic the estimates regarding when students should be ready or when testing should provide us with an accurate picture of student learning. We have yet to mention issues like student mobility, school attendance, and prior educational background which are all very important in determining readiness to participate in large-scale assessments; especially when those assessments count educationally either for or against the children who take them.

Access to Content: The Need for Quality Academic Support Programs

Rivera and Vincent (1996) and Zehler et al. (1994) raise the issue of LEP student access to academic content and the related issue of program quality. Even if LEP students have been in American schools three years or longer, that information alone tells us little about the length of time they have had exposure to academic content through bilingual instruction, content-based ESL classrooms or adequately adapted or supported mainstream classrooms. Of course, clearly distinguishing between good and bad programs is the principal reason inclusive assessment and accountability for LEP students is so important. On the other hand, holding individual children accountable for content that was never adequately taught is hardly reasonable and certainly not the cure for the already high minority student dropout rate. (LaCelle-Peterson & Rivera, 1994; McDill et al., 1985)

Access to content is almost universally cited as an issue of importance, but few authors treating the topic of large-scale assessment of LEP students offer a definition of quality program support, or clearly explain how long students might be expected to need such support. Seemingly not addressed by the literature is the possibility that early and inappropriate inclusion of LEP students in large-scale assessments might actually have a negative, not positive, effect on the programs that serve LEP students. Bilingual educators have argued for years that late-exit programs, which teach both English and the child's primary language, work better than the more rapid, intensive English instruction which often sacrifices high quality academic content for a quicker transition to English. Gandara & Merino (1993), in their case study of bilingual/ESL programs, maintain that scores on assessments taken in English after one and one-half years of English exposure may well be superior for early-exit, intensive, remedial English programs. Content-based programs that use both English and the primary language may require three years or longer to demonstrate their superiority to the "quick fix" solution.

The type of assessment students take for accountability is another important issue. Traditional norm referenced testing, still used by some states, is considered particularly prone to issues of cultural and linguistic bias. Criterion referenced testing, emerging as the favorite in the standards-based era, should offer advantages, but only if students have had an adequate opportunity to master the content being covered. Some states now include performance assessment items with large-scale assessments, but these questions can present even more linguistic challenges than multiple choice. (Anderson et al., 1997a)

High Stakes Consequences

Large-scale assessments increasingly come with high stakes consequences for the students who take them. Most state assessment programs began without such consequences but state legislatures around the nation have and continue to pass laws that require retention-in-grade (often known as “no-social promotion” legislation) and mandatory achievement levels in high school graduation tests. Ironically, states have tended to maintain liberal exemption policies for LEP students on tests that have little or no state mandated consequences for failure; on the more recently enacted high stakes assessments, the trend, Anderson et al. (1997a) points out, is to permit few, if any, exemptions, and often few, if any, testing accommodations.

This is precisely the use of test results that the literature overwhelmingly warns us against. Messick (1994) refers to this as “consequential validity,” or the unintended consequences that come from using test results of dubious validity for important educational decisions. Olsen and Goldstein (1997) are even more to the point: “Until the psychometric issues underlying these new assessments have been addressed, and until mechanisms to ensure opportunities to learn have been fully implemented, these assessments should not be used in high stakes situations.” (P.50)

Use of Accommodations When Testing LEP Students with Large-Scale, Academic Assessments

Testing accommodations is an area where it is worthwhile to mark a contrast with special education. The number and variety of issues facing students with disabilities requires far greater accommodations options than those that might benefit LEP students. Nonetheless, the literature suggests that selective use of testing accommodations that do not significantly affect the validity of large-scale assessments should be encouraged for LEP students as this may minimize to some degree the linguistic barriers and thus provide a truer picture of student content knowledge. (Olsen & Goldstein, 1997) The literature lists the most commonly permitted accommodations for LEP students as: additional time, bilingual dictionaries, translating test directions into the native language, and small group testing administration. (Anderson et al., 1997a)

The literature identifies three problems that exist with the use of accommodations for LEP students: a) very little research has been done to indicate which accommodations will benefit LEP students; b) educators have little evidence to conclude that a particular accommodation does or does not affect the validity of the assessment, and c) while most states are willing to accommodate students with disabilities, many appear reluctant to afford LEP students even a few of the most basic accommodations like additional time, dictionaries, and small group testing administration. (NCBE, 1997)

Problem “a” may explain, in part, the reluctance to allow LEP students accommodations when taking standardized tests. However, there is less reluctance to allow accommodations for students with disabilities even though the research is also inconclusive. This reluctance to allow accommodations for LEP students becomes more evident when high stakes penalties are attached to those same assessments. This is because testmakers demand higher levels of reliability from assessments as consequences are associated with their use.

Another reason for the reluctance to accommodate LEP students may be the general perception that students should not need more than a year or two to catch up with their English speaking peers; a perception we may actually perpetuate further by recommending testing after one and one-half years in American schools. Again, this reluctance seems greater as the stakes

get higher. Anderson et al. state: “We found that state exit exam policies almost always were applied equally to LEP and regular education students.” (p.9)

Some states do allow limited accommodations for LEP students like those mentioned earlier. But again, researchers are still unsure to what degree individual accommodations like providing additional time serve to level the playing field. New York is one of the few states that goes so far as to say students may take the assessments in their native languages (the top five in the state), or have an interpreter present during the entire testing period to translate individual questions. (Personal communication with New York State Educational Agency, February, 1999.)

The issue of translating the test as a testing accommodation brings with it its own dilemmas. Translations are difficult to do as students speak multiple languages and multiple dialects within those languages. Issues of insuring comparability between the original and translated versions present an almost insurmountable obstacle to assessment experts. Translations also assume that students have either had instruction in their native language within the U.S., or had it before they arrived here. (NCBE, 1997; Olsen & Goldstein, 1997)

Alternate Assessment of LEP Students

When educators think of alternatives to traditional, large-scale assessments, classroom-based portfolio or performance assessment is usually mentioned. Anderson et al. (1997a/b) point out that these assessments have several drawbacks, however. First, they may still be very difficult linguistically, in some cases more difficult than traditional test questions. Scores from alternate assessments, as they are now called, are not easily equated with those from more traditional, large-scale assessments. When assessments are scored at the local school level by various “raters,” issues of inter-rater reliability lead to serious questions of comparability between scores. When these issues are considered in light of the increasingly popular high stakes purposes to which large-scale assessments are being used, it is easy to see why many educators do not consider alternate assessment as a viable option.

Notwithstanding the drawbacks, use of alternate assessment is becoming popular, particularly with students with disabilities in the more severe categories where the individualized educational plan (IEP) goals may differ greatly from the mainstream curriculum. Some states have adopted alternate assessments that are more functional than academic in an attempt to meet the needs of students with cognitive disabilities, autism, or other severe conditions affecting learning. These assessments include students within the accountability framework, albeit in ways which are largely incomparable with standardized assessment. A similar movement towards alternate assessment has not taken place for LEP students, again perhaps because the assumption has been that they can and should participate in the regular assessment within a relatively short period of time. In the high stakes testing era, educators might be wise to rethink the validity and fairness of that assumption and ask whether classroom-based, alternate assessments, while not comparable to large-scale assessments, could fulfill a local accountability role as students gain the English proficiency and level of acculturation necessary to succeed in the regular assessment system. (McDonnell et al., 1997; Anderson et al., 1997a)

Rivera and Vincent (1996) state that alternate assessment should flow naturally from the language assistance program, should reduce the linguistic load but reflect the same high content standards of the mainstream. This is a very different type of assessment than the commercially available language proficiency assessments, largely devoid of academic content,

that are typically used to measure LEP student growth while schools wait for students to reach higher English proficiency levels.

Conclusions

Most of the current literature advocates that LEP students be included in large-scale assessment as soon as possible. Whereas, traditionally, educators would exclude students when in doubt as to the LEP student's readiness to participate, the trend today is, if anything, to err on the side of inclusion. This is probably the best approach when the purpose of testing is to monitor student progress or program performance over time. Educators appear to be unanimous in their belief that LEP students, and the programs that serve them, have been left out of the accountability system far too long. Within such an accountability framework, the results of LEP assessments should be examined over time and the accountability yardstick must be full academic parity, not merely quick gains in English. Otherwise, even the broad accountability function of large-scale assessment may not result in improvements that are congruent with best practice.

Increasingly, however, the purposes which large-scale assessments serve include more than merely monitoring progress over time. Consequences for individual children and the programs that serve them may be based on a single point-in-time assessment. Adding new consequences to the same or similar testing instrument should give us cause to consider the validity of making important educational decisions based on the results of large-scale assessments. This is particularly true when those assessments are offered with limited accommodations or alternatives, and scores reflect one point in the LEP student's academic path.

When LEP students participate in large-scale assessments at intermediate levels of English proficiency, educators need to give serious considerations to allowing accommodations. While certainly there is much still to learn about their use, accommodations offer at least limited promise to level the assessment playing field not only for students with disabilities, with whom their use is already common, but with LEP students at intermediate levels of English proficiency.

Alternate assessments conducted at the local, classroom level, while not without problems, offer a measure of accountability for academic progress for beginning to intermediate level LEP students that has been absent to date. These alternate assessments, when grounded in common academic performance standards, may provide teachers of LEP students with a tool for aligning curriculum and improving instructional monitoring of individual students and the programs that serve them.

By increasing the educational stakes and insisting that all children participate equally, we may, in the name of accountability, draw invalid conclusions about LEP students and the language assistance programs that serve them. When large-scale assessments serve high stakes purposes, we need to reassess the standards for validity and the consequences that follow.

Bibliography

Anderson, C., Wagner, C., & Tillman, M. (1997a). *A review of the literature on students with limited English proficiency and assessment.* Minneapolis, MN: National Center on Educational Outcomes.

Anderson, C., Wagner, C., & Tillman, M. (1997b). *High school graduation requirements in the U.S. for students with limited English proficiency.* Minneapolis, MN: National Center on Educational Outcomes.

Anstrom, K. & Kindler, A. (1996). *Federal policy, legislation and education reform: The promise and the challenge for language minority students*. National Clearinghouse for Bilingual Education, Resource collection series no. 5. Washington, DC: George Washington University.

August, D. & Hakuta, K. (Eds.) (1997). *Improving schooling for language-minority children: A research agenda.* National Research Council, Washington, DC: National Academy Press.

Author. (1997). Individuals with Disabilities Education Act (IDEA) Amendments of 1997. Sec. 101 - 203, 1-107. ftp://ftp.loc.gov/pub/thomas/c105/h5.enr.txt.

Author. (1996). *Title I, Part A, Policy Guidance: Improving basic programs operated by local educational agencies.* U.S. Department of Education, Office of Elementary and Secondary Education, Compensatory Education Programs: Washington, DC.

Butler, F. & Stevens, R. (1997). *Accommodation strategies for English language learners on large-scale assessments: Student characteristics and other considerations.* CSE Technical Report 448. University of California, Los Angeles, CA: Center for the Study of Evaluation.

Gandara, P. & Merino, B. (1993). Measuring the outcome of LEP programs: Test scores, exit rates and other mythological data. *Educational Evaluation and Policy Analysis,* 15 (3), 320-338.

Gottlieb, M. (in press). *Standards-Based alternate assessment for limited English proficient students: A guide for Wisconsin Educators.* Madison, WI: Wisconsin Department of Public Instruction.

LaCelle-Peterson, M. & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review,* 64 (1), 55-75.

McDill, E. L., Natriello, G., & Pallas, A. (1985). Raising standards and retaining students: The impact of the reform recommendations on potential dropouts. *Review of Educational Research*, 55 (4), 415-433.

McDonnell, L.M., McLaughlin, M.J. & Morrison, P. (Eds.) (1997). *Educating one & all: Students with disabilities and standards-based reform.* National Research Council, Washington, DC: National Academy Press.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23(2), 13-23.

Olsen, J. & Goldstein, A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress.* Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement.

USDE. (1997). *High stakes assessment: A research agenda for English language learners.* Symposium summary. National Clearinghouse for Bilingual Education. Washington, DC: George Washington University.

Zehler, A., Hopstock, P., Fleishman, H., & Greniuk, C. (1994). *An examination of assessment of limited English proficient students.* Arlington, VA: Special Issues Analysis Center.

[1] While the term English Language Learner (ELL) is preferred by many language minority educators today, Limited English Proficient (LEP) is still the designation found in federal and state legislation and policy guidance.

[2] While special educators estimate that between 10-20% of all students with disabilities will take alternate assessments, no similar national level estimates exist for LEP students as states differ widely regarding how long students are allowed exemptions from standardized tests and what, if any, alternate assessments are required in the interim.